J. Clin. Chem. Clin. Biochem.
Vol. 23, 1985, pp. 739–752

The Classification of Subjects with Joint Complaints on Incomplete Biochemical and Haematological Datasets

By *H. M. J. Goldschmidt, J. den Hartog*[1]), *J. F. Leijten*

Department of Clinical Chemistry and Haematology, Maria Hospital, Tilburg, The Netherlands

D. Coomans and *D. L. Massart*

Pharmaceutic Institute, Free University of Brussels, Brussels, Belgium

(Received October 4, 1982/February 25, 1985)

Summary: We performed a retrospective study on 163 subjects suffering from rheumatic fever (16), rheumatoid arthritis (36), lupus erythematosus (17), gout (21), arthrosis (50) and osteomyelitis (23).

The number of variables evaluated was 39. These were all of a general biochemical and haematological nature. A feature reduction resulted in sixteen variables that matched well with those known from the literature.

Linear discriminant analysis yielded poor results in classifying the six disease categories (with 18 variables 61.8%).

A reduction to three disease categories improved the classification results remarkably. This, and the excellent discriminating power between patients and the reference group, shows that the selected variables are illustrative only for general clinical pictures, such as infection, and not for the desired differential diagnosis.

Klassifizierung von Personen mit Gelenkbeschwerden aufgrund unvollständiger biochemischer und hämatologischer Datensätze

Zusammenfassung: Wir führten eine retrospektive Studie an 163 Personen, die an rheumatischem Fieber (16), rheumatischer Arthritis (36), Lupus erythematodes (17), Gicht (21), Arthrose (50) und Osteomyelitis (23) erkrankt waren, durch.

Die Anzahl der ausgewerteten Variablen betrug 39. Diese waren alle allgemeiner biochemischer und hämatologischer Natur. Eine Verminderung der Merkmale ergab 16 Variable, die gut mit den aus der Literatur bekannten übereinstimmten.

Lineare Diskriminanzanalyse ergab für die Klassifizierung der sechs Krankheitskategorien schlechte Ergebnisse (61,8% mit 18 Variablen). Eine Verminderung auf drei Krankheitskategorien verbesserte die Klassifizierungsergebnisse bemerkenswert. Dies und die hervorragende Diskriminierung zwischen Patienten und Referenzkollektiv zeigt, daß die gewählten Variablen nur für allgemeine klinische Bilder wie Infektion bezeichnend sind, nicht aber für die erstrebten Differentialdiagnosen.

[1]) Present address: Pharmacia Nederland B.V., Ohmweg 12, NL-3442 AA Woerden.

Introduction

In recent years an increasing number of articles have been written on the use of multivariate analysis for clinical chemical problems. The review of *Goldberg & Ellis* (1) gives many examples in the different application areas. It shows that multivariate analysis can be applied succesfully for detecting common features, for instance, in the search for risk factors. It also demontrates the possibilities of multivariate analysis in data reduction, by quantitating the extra information provided by an additional analysis. Finally, it reveals various examples of how to classify patients in different disease categories. This study deals with the classification of patients with joint complaints. These subjects were put into six classes based upon their complaints, radio-graphic findings and external appearance.

Usually many physical, chemical and haematological data are gathered and this raised the following questions:

a) are physical, chemical and haematological data useful in discriminating between these patients and is the complete dataset needed or is part of the data sufficient?

b) are these data or a subset operative in the distinction between a normal population and patients with joint disease?

c) which variables are pathognomic for which disease?

The answers to these questions would lead to a reduction in the number and the variety of tests. Any good strategy will reveal the importance of serological tests but, since we are only interested in the relevance of higher order variables, we excluded serological tests from our dataset. Also, when dealing with descriptive variables, it is not surprising that much potential data for each patient is not obtained.

An earlier study of *Wilding* et al. (2) was focused on changes in chemical and haematological data caused by drugs or due to the disease activity, but solely in cases of rheumatoid arthritis. The present paper deals with more variables, a broader scale of diseases and the problem of missing data.

Our purpose was not only to distinguish the patient groups from a group of apparently healthy individuals, but also to separate different disease categories from each other.

This study deals with a set of patients (objects), each of which generates the values of a certain number of test results (variables or features). Therefore a multivariate approach is necessary. Classification of objects and grouping of variables are the main purposes of this work.

The first problem is to reduce the number of variables to a manageable set. Assuming that tests not usually asked for are less meaningful, we left out all those variables lacking in 25% or more of the patients. The remaining were ordered and reduced, using two feature selection procedures: variance weighting and the multivariate F-ratio. We used the selected set of variables in a linear discriminant approach to the classifiation of the patients in the different disease categories.

Materials

The study involves six disease categories as shown in table 1. The reference group is a group of apparently healthy blood donors, varying adequately in sex and age.

Tab. 1. Outline of selected groups of patients.

Group no.		♂	♀	Total
1	Rheumatic fever	8	8	16
2	Rheumatoid arthritis	18	18	36
3	Lupus erythematosus	3	14	17
4	Gout	14	7	21
5	Osteoarthrosis	21	29	50
6	Osteomyelitis	11	12	23
		75	88	163
7	Healthy blood donors	57	29	86

In the record offices of the St. Elisabeth and Maria Hospitals in Tilburg, The Netherlands, we looked for those patients whose main disease state corresponded to one of the six mentioned disease classes. All patients were classified according to the ICD.9.CM index. An independent physician read the medical records and checked the diagnosis. In addition, we interviewed the second consultant with regard to the reliability of the diagnosis. From the responses of these three (the first attending physician, the independent and the second consultant) we concluded that the diagnosis for the patients in question were all well established.

Variables

From the medical records of the 163 patients, all those data were taken which had been obtained within 24 hours (48 hours in weekends) after admission. There proved to be 150 different variables. The 24 hour time limit was chosen to minimize time, therapeutic and drug effects.

It seemed reasonable to suppose, and it was admitted by the consultant physicians, that data taken only incidentally were not gathered in relation to the joint complaints. As a first selection criterion, we therefore limited the variables to those found in at least 120 patients (= 75%). By doing this the 39 variables of table 2 remained. These 39 variables correspond with the variables normally gathered in an admission and screening profile.

To determine haematocrit, haemoglobin, erythrocytes, leukocytes and platelets a Hemalog 8 (Technicon Instruments B.V., Gorinchem, The Netherlands) was used. Differential counts (neutrophil segmented granulocytes, neutrophil band form granulocytes, lymphocytes, monocytes, eosinophil granulocytes, basophil granulocytes) were determined by manual eye microscopy. The Erythrocytes Sedimentation Rate was calculated after 1 hour of standing in a 20 cm sedimentation tube. The SMA-C (Sequential Multi Analyzer plus Computer, Technicon Instruments B.V., Gorinchem, The Netherlands) was used in the assay of the analytes in serum. Standard bicarbonate was determined using a Radiometer pH meter (Model PHM, Radiometer Corp. Copenhagen, Denmark) according to the method of *Jørgensen* (3). Conjugated bilirubin in serum was determined on an AutoAnalyzer II (Technicon Instruments B.V., Gorinchem, the Netherlands) using the method of *Borst* (4).

The thymol turbidity analysis was performed in serum by the method of *Kingsley & Getchell* (5). Pulse rate, systolic and diastolic blood pressure, body weight, body temperature and age were all determined on admission as part of the routine physical examination.

It should be noted again that the rheuma serology data were not included, because their diagnostic efficiency is known and our object was to study the value of the other tests in relation to the kind and severity of the disease.

Table 2 gives the range and the mean of the different variables for the six disease classes and the reference group. The numerical value of each test result for the appropriate number of variables was used to calculate, by linear discriminant analysis, the allocation of each patient to a particular disease category. By feature reduction the total number of variables necessary to discriminate all the disease categories from the reference population was determined to be 10. The decision to allocate any patient to a disease or non-disease category is based on the statistical analysis of all included variables and not on the absolute value of any individual test result.

Statistical Methods

We were left with 39 variables. In relation to the number of patients in the various disease classes, this number is too large for multivariate techniques like linear discriminant analysis. We therefore introduced an intermediate step for further reduction of the number of variables, which uses a combination of two feature selection procedures: feature selection by variance weighting and stepwise feature selection on the basis of the multivariate F-ratio.

Feature selection by variance weighting (6, 7) is a statistical technique in which each variable is weighted on the basis of its individual importance in the differentiation between each pair of diagnostic classes (no correlation is taken into account). The variance weight for two classes is the ratio of the interclass variance of the two classes to the intraclass variances of these classes (univariate F-ratio). The higher the variance weight, the higher the individual importance of the tests. This procedure was performed using the software package ARTHUR (7).

Stepwise feature selection was performed on the basis of the multivariate F-ratio (8). The method was applied using the software package SPSS (9). In the stepwise procedure, combinations instead of individual variables are considered. Initially, the single test which has the best value for the selection criterion is chosen; in the present case it is the F-ratio (the univariate one in the first step of the procedure and the multivariate one in the further steps) for the separation of the centroids of the diagnostic classes. This initial variable is then paired sequentially with each of the other available variables and the F-ratio is computed again. The variable which produces the most significant increase of the F-ratio is selected as the second variable. The procedure continues in this way until all variables are included or the best variable in a given step of the procedure does not permit a significant increase of the multivariate F-ratio. For the multivariate F-ratio, correlations between the variables are taken into account.

In multivariate techniques there is a relation between the maximum permissible variables and the number of patients in a disease class, because too many variables will result in unstable estimates of the differences between the class centroids. Since the smallest diagnostic class contains only 16 patients (rheumatic fever) not more than 16 laboratory tests were selected in every paired weighting. the number of variables was thus reduced to 16.

The discriminatory performance of the laboratory tests was further investigated by means of linear discriminant analysis (8). In general terms, linear discriminant analysis distinguishes diagnostic classes on the basis of a set of linear functions of the variables. The weight coefficients in these functions are calculated in such a way that the ratio of between-class to within-class variation is maximized, considering also the correlation between the variables. Linear discriminant analysis was performed using the program package SPSS. The method was not applied to pairs of classes, as in the selection procedures, but directly to the seven diagnostic classes. This was done in order to obtain a better estimated pooled variance-covariance matrix with more patients, especially when a large number of variables are being used (for instance 20 laboratory tests).

Results

Feature selection for discriminating between different disease categories

With the described two feature selection methods the data in the six disease classes were compared mutually, one class with another. In each of the resulting 15 binary discriminating problems, only those variables were used which were present in at least 75% of the patients involved in the particular paired classification problem. In this way we made use, in the comparison between gout and arthrosis, of only 10 variables, whereas in the comparison between rheumatic fever and lupus erythematosus, 33 variables were used. Tables 3 and 4 give the results.

For reasons mentioned above, only the six variables with the highest score are given, and they are ranked in order of decreasing value. At this stage we were left with 25 variables, which are mentioned at least once in tables 3 and 4.

Tab. 2. The matrix displays in each cell the mean and range (the lowest and highest value) for each variable in each group.

	Variables	Disease categories							Units
		Rheumatic fever	Rheumatoid arthritis	Lupus erythematosus	Gout	Arthrosis	Osteomyelitis	Reference group	
1	Haematocrit	0.46–0.47 0.462	0.34–0.51 0.397	0.29–0.45 0.396	0.37–0.50 0.443	0.31–0.48 0.408	0.36–0.49 0.409	0.35–0.55 0.436	1
2	Haemoglobin	6.2–10.6 8.86	6.3–10.6 8.36	5.2–9.7 8.49	5.4–12.9 9.62	7.0–10.8 9.01	7.2–10.4 8.20	6.8–12.3 9.52	mmol/l
3	Erythrocytes	4.1–5.4 4.74	2.8–5.3 4.35	3.7–5.2 4.37	2.7–6.3 4.84	3.7–6.4 4.66	3.8–5.4 4.47	4.0–6.4 4.83	10^{12}/l
4	Leukocytes	4.1–47.0 11.41	2.4–17.9 7.06	3.0–10.4 5.90	4.3–13.3 7.88	3.5–14.9 6.95	3.3–21.6 7.62	3.6–10.8 6.50	10^9/l
5	Platelets	303–460 364.0	94–506 265.6	140–426 249.0	210–295 251.2	1–400 8.4	112–546 341.3	163–371 243.8	10^9/l
6	Serum sodium	135–144 140.5	132–149 141.2	130–150 140.8	138–147 142.6	130–146 140.8	136–145 141.0	137–144 140.3	mmol/l
7	Serum potassium	4.0–4.7 4.34	3.3–5.1 4.35	3.7–5.4 4.40	3.5–5.0 4.20	3.0–5.3 4.26	2.6–5.2 4.36	3.0–4.9 3.96	mmol/l
8	Serum chloride	99–107 102.9	93–108 100.8	96–108 102.2	94–109 103.3	92–110 100.8	94–110 101.0	93–106 99.5	mmol/l
9	Standard bicarbonate	24.3–25.2 24.75	18.4–33.1 24.67	22.3–27.8 24.65	22.7–30.3 25.74	23.0–28.0 25.42	16.9–28.8 24.48	22.0–34.0 27.63	mmol/l
10	Serum creatinine	71–88 79.6	55–186 83.2	50–97 73.0	87–124 109.7	53–126 77.6	35–232 92.1	64–142 95.5	µmol/l
11	Serum uric acid	0.17–0.44 0.288	0.13–0.58 0.297	0.20–0.45 0.286	0.11–0.72 0.434	0.14–0.72 0.352	0.12–0.70 0.289	0.17–0.55 0.330	mmol/l
12	Serum calcium	2.3–2.6 2.42	2.0–2.7 2.37	2.0–2.4 2.29	1.9–3.0 2.44	2.1–2.6 2.41	2.1–2.6 2.36	2.3–2.7 2.50	mmol/l
13	Serum inorganic phosphate	1.00–1.81 1.305	0.32–1.62 1.001	0.68–1.39 1.046	0.74–1.65 1.104	0.48–1.38 .969	0.36–1.69 1.142	0.52–1.47 1.083	mmol/l
14	Serum glucose	3.2–9.7 6.60	3.6–17.2 6.71	3.6–9.7 5.66	4.5–14.5 7.81	3.5–12.6 6.71	5.1–16.5 8.81	2.9–11.2 5.37	mmol/l
15	Serum total protein	67–78 72.5	59–100 73.0	61–89 74.7	63–84 73.0	61–88 72.8	61–78 69.7	66–81 72.7	g/l
16	Serum albumin	28–51 39.4	21–55 39.3	26–48 38.1	24–44 37.5	27–50 42.6	31–45 38.3	39–51 46.0	g/l
17	Serum total bilirubin	2.6–14.5 7.14	2.7–29.1 7.70	1.9–20.5 8.01	1.7–17.4 9.39	3.9–55.1 10.76	3.4–17.3 7.09	3.0–22.0 8.59	µmol/l
18	Serum alkaline phosphatase	50–246 124.9	44–260 102.6	52–280 105.7	10–195 69.6	37–350 83.3	9–228 118.7	36–189 76.3	U/l
19	Serum lactate dehydrogenase	120–341 188.1	119–374 193.8	145–564 225.3	119–190 151.6	87–323 182.7	130–390 206.9	104–278 184.0	U/l

Tab. 2. Continued

20	Serum aspartate aminotransferase	10–85 42.6	12–118 27.6	16–163 46.2	10–63 31.1	13–253 32.8	7–74 25.9	6–49 17.8	U/l
21	Serum total cholesterol	3.9–6.9 5.13	3.2–7.8 5.38	3.2–8.5 5.62	4.4–9.8 6.31	3.7–8.7 5.93	3.4–6.2 4.51	4.5–10.0 6.91	mmol/l
22	Erythrocytes sedimentation rate	3–112 41.7	6–115 39.0	12–129 41.0	3–150 32.5	2–90 18.4	19–84 49.3	1–39 6.7	mm/1 h
23	Serum conjugated bilirubin	1.7–3.0 2.35	1.7–14.4 3.34	2.2–19.5 6.71	2.6–3.2 2.91	1.7–8.9 3.20	2.0–5.0 2.99	0.5–5.0 2.0	µmol/l
24	Serum urea	3.7–7.2 5.71	2.5–14.9 6.40	3.0–10.0 5.74	2.2–12.2 6.10	3.2–13.5 6.52	2.7–25.0 6.35	1.8–9.4 5.47	mmol/l
25	Pulse rate	64–100 86.5	50–116 80.5	64–100 81.1	56–124 77.7	60–104 79.3	60–104 80.0	53–91 72.0	min^{-1}
26	Systolic blood pressure	90–180 139.4	115–230 154.8	90–220 149.1	110–220 157.5	115–220 166.1	85–280 144.2	110–190 139.9	mmHg
27	Diastolic blood pressure	65–110 91.7	70–110 90.6	60–110 85.6	60–140 97.5	70–120 95.9	45–130 80.8	60–120 89.2	mmHg
28	Serum alanine aminotransferase	7–110 58.5	5–133 24.3	4–183 34.0	6–21 14.2	7–37 17.9	9–128 37.4	15–54 34.6	U/l
29	Serum creatine phosphokinase	72–200 136.0	26–424 137.8	51–706 150.3	79–336 190.8	64–268 175.2	90–163 135.7	29–565 107.9	U/l
30	Serum thymol turbidity	1–5 3.0	1–6 3.5	1–15 4.1	1–5 3.0	1–18 6.5	1–5 3.0	1–4 2.0	U
31	Weight	12.3–86.0 62.00	42.2–91.6 63.44	42.5–74.6 63.11	56.5–101.0 76.51	52.0–105.5 72.78	9.6–69.0 45.79	53.0–94.0 73.56	kg
32	Neutrophil segmented granulocytes	26–80 64.4	36–88 68.7	52–86 70.1	41–79 64.6	46–87 65.9	31–88 65.1	44–69 56.2	/100
33	Neutrophil band form granulocytes	0–11 4.2	0–7 0.9	0–10 2.1	0–7 2.6	0–11 1.1	0–8 1.7	0–8 1.3	/100
34	Lymphocytes	9–70 26.9	4–61 25.1	10–41 23.6	13–51 27.7	4–49 27.7	1–64 26.7	22–42 32.8	/100
35	Monocytes	1–9 3.8	0–9 3.4	0–7 2.5	0–9 3.4	0–10 3.0	1–6 3.1	0–10 5.0	/100
36	Eosinophil granulocytes	0–3 0.4	0–6 1.8	0–6 1.4	0–8 1.8	0–9 2.0	0–3 1.0	0–9 2.4	/100
37	Basophil granulocytes	0–1 0.1	0–2 0.1	0–1 0.1	0–2 0.2	0–1 0.1	0–1 0.1	0–2 0.6	/100
38	Body temperature	37.6–38.9 38.30	35.6–38.4 37.36	36.1–39.6 37.91	35.8–39.1 37.48	35.4–38.0 36.68	36.5–40.8 38.09	36.2–37.6 36.90	°C
39	Age	0–99 32.9	6–90 57.6	18–73 52.5	17–82 53.2	24–85 61.7	1–95 35.9	19–63 36.0	years

Tab. 3. No correlation correction, binary comparisons with the variance weight criterion.

	Rheumatic fever	Rheumatoid arthritis	Lupus erythematosus	Gout	Arthrosis	Osteomyelitis
Rheumatic fever		Neutrophil band form granulocytes	Serum inorganic phosphate	Serum total cholesterol	Serum inorganic phosphate	Serum total cholesterol
		Serum inorganic phosphate	Serum calcium	Serum uric acid	Neutrophil band form granulocytes	Eosinophil granulocytes
		Eosinophil granulocytes	Serum glucose	Age	Age	Serum total protein
		Age	Leukocytes	Eosinophil granulocytes	Eosinophil granulocytes	Haemoglobin
		Leukocytes	Eosinophil granulocytes	Neutrophil band form granulocytes	Erythrocytes sedimentation rate	Serum inorganic phosphate
		Haemoglobin	Neutrophil band form granulocytes	Leukocytes	Serum alkaline phosphatase	Leukocytes
Rheumatoid arthritis			Serum aspartate aminotransferase	Haemoglobin	Erythrocytes sedimentation rate	Serum total cholesterol
			Serum calcium	Neutrophil band form granulocytes	Haemoglobin	Age
			Neutrophil band form granulocytes	Serum uric acid	Serum uric acid	Eosinophil granulocytes
			Serum creatinine	Serum total cholesterol	Serum albumin	Serum inorganic phosphate
			Leukocytes	Neutrophil segmented granulocytes	Serum total cholesterol	Serum total protein
			Serum lactate dehydrogenase	Leukocytes	Serum alkaline phosphatase	Neutrophil band form granulocytes
Lupus erythematosus				Serum uric acid	Serum calcium	Serum total cholesterol
				Haemoglobin	Erythrocytes sedimentation rate	Serum glucose
				Leukocytes	Serum albumin	Serum total protein
				Diastolic blood pressure	Serum glucose	Serum aspartate aminotransferase
				Neutrophil segmented granulocytes	Serum uric acid	Age
				Serum potassium	Serum lactate dehydrogenase	Leukocytes

Tab. 3. Continued

Gout	Serum sodium	Serum total cholesterol
	Serum uric acid	Haemoglobin
	Serum chloride	Serum uric acid
	Haemoglobin	Age
	Erythrocytes sedimentation rate	Erythrocytes sedimentation rate
	Age	Eosinophil granulocytes
Arthrosis		Erythrocytes sedimentation rate
		Serum total cholesterol
		Serum albumin
		Age
		Haemoglobin
		Serum inorganic phosphate

Now each of the six variables was given points according to its ranking order in tables 3 and 4. The points in each of the two comparisons were summed, and the three variables with the highest score for each binary comparison are given in table 5. The number of variables used in every comparison are given in table 5 at the bottom right of each square.

After this operation there now remained 16 variables that are mentioned at least once. Textbooks of pathology (10–12) also mention laboratory findings that more or less regularly accompany the various diseases. A compilation of these observations shows that the following are usually proposed as a monitor:

Rheumatic fever:

body temperature, age, leukocytes, erythrocytes sedimentation rate, haemoglobin.

Rheumatoid arthritis:

haemoglobin, erythrocytes sedimentation rate, sex, leukocytes, platelets.

Lupus erythematosus:

haemoglobin, age, platelets, serum albumin, erythrocytes sedimentation rate, leukocytes, serum total protein.

Gout:

sex, serum uric acid, serum glucose, systolic blood pressure, diastolic blood pressure.

Arthrosis:

age, sex.

Osteomyelitis:

neutrophil segmented granulocytes, leukocytes, erythrocytes sedimentation rate.

Table 6 shows the 16 variables, found by our operation and mentioned in table 5, in relation to the 13 'textbook variables', and shows that there is a strong similarity between the variables we calculated and the findings mentioned in textbooks.

Tab. 4. Partial correlation correction, binary comparisons taking into account that the variables were selected in order of maximizing the smallest F-ratio between pairs of groups.

	Rheumatic fever	Rheumatoid arthritis	Lupus erythematosus	Gout	Arthrosis	Osteomyelitis
Rheumatic fever		Neutrophil band form granulocytes	Neutrophil band form granulocytes	Serum total cholesterol	Serum inorganic phosphate	Haemoglobin
		Serum inorganic phosphate	Serum glucose	Neutrophil band form granulocytes	Erythrocytes sedimentation rate	Serum total cholesterol
		Haemoglobin	Serum inorganic phosphate	Serum uric acid	Neutrophil band form granulocytes	Serum inorganic phosphate
		Lymphocytes	Haemoglobin	Lymphocytes	Leukocytes	Eosinophil granulocytes
		Erythrocytes sedimentation rate	Serum calcium	Eosinophil granulocytes	Haemoglobin	Monocytes
		Eosinophil granulocytes	Serum total protein	Age	Serum calcium	Leukocytes
Rheumatoid arthritis			Serum calcium	Haemoglobin	Age	Serum total cholesterol
			Serum glucose	Monocytes	Serum uric acid	Serum uric acid
			Serum total protein	Neutrophil band form granulocytes	Serum lactate dehydrogenase	Neutrophil band form granulocytes
			Leukocytes	Erythrocytes sedimentation rate	Neutrophil band form granulocytes	Erythrocytes sedimentation rate
			Neutrophil segmented granulocytes	Serum uric acid	Lymphocytes	Age
			Serum total cholesterol	Age	Serum glucose	Haemoglobin
Lupus erythematosus				Haemoglobin	Erythrocytes sedimentation rate	Age
				Serum chloride	Age	Serum glucose
				Serum potassium	Serum calcium	Serum calcium
				Leukocytes	Serum inorganic phosphate	Serum total protein
				Lymphocytes	Serum potassium	Erythrocytes sedimentation rate
				Serum sodium	Serum chloride	Serum total cholesterol

Tab. 4. Continued.

Gout	Haemoglobin	Haemoglobin
	Erythrocytes sedimentation rate	Neutrophil band form granulocytes
	Serum chloride	Serum total cholesterol
	Serum potassium	Leukocytes
	Serum uric acid	Serum uric acid
	Serum total cholesterol	Erythrocytes sedimentation rate
Arthrosis		Age
		Erythrocytes sedimentation rate
		Serum glucose
		Serum uric acid
		Serum albumin
		Haematocrit

Feature reduction while discriminating between different disease categories and a reference group

After the selection of the most meaningful variables, the classification of patients was studied using linear discriminant analysis. This method is well known in clinical chemistry (1, 2).

When a patient is characterized by 10 variables, the patient may be represented as a point in 10-dimensional space. Patients situated near to each other have similar patterns and, when the variables are sufficiently relevant, suffer from the same disease (or are both healthy).

Unfortunately, one cannot directly view a 10-dimensional space, but methods are available to represent this hyperspace in an optimal way in only two dimensions. The linear discriminant functions developed in linear discriminant analysis permit this.

Figure 1 a gives such a discriminant plot for the utilization of 20 variables; 1 b is the same plot but includes only 10 variables, and 1 c shows the plot when only 9 variables are used. They demonstrate, as denoted in the legend to these figures, that it is possible to distinguish more or less between the seven classes, because about 60% of the subjects are classified correctly. These plots reveal that the reference group can be separated easily but only when more than 9 variables are included. When the variable, age, is eliminated in going from 10 to 9 variables, the distinction disappears.

Variables pathognomonic for each disease category

In this study the number of patients is so small that it is impossible to split the database to create an independent test set. When using the learning set as test set, in the linear discriminant analysis, over-optimistic results can be obtained. The leave-one-out procedure is the solution in situations like these.

Statistically this is comparable with an independent dataset. A discriminant analysis is performed on the total number of patients minus one. Each time another one is omitted and is classified by means of his discriminant scores in the accompanying analysis. The discrimination between the six disease categories on the basis of chemical and haematological variables is poor. When using the leave-one-out procedure, only 40% of the patients are classified correctly with 18 variables, while with 10 variables, only 31% are classified correctly.

Tab. 5. Joined variable matrix based on binary statistical comparisons with indication of the total number of variables used for selection in tables 2 and 3.

Rheumatic fever	Rheumatoid arthritis	Lupus erythematosus	Gout	Arthrosis	Osteomyelitis
Rheumatic fever	Neutrophil band form granulocytes	Serum inorganic phosphate	Serum total cholesterol	Serum inorganic phosphate	Serum total cholesterol
	Serum inorganic phosphate	Serum glucose	Serum uric acid	Neutrophil band form granulocytes	Haemoglobin
	Haemoglobin	Neutrophil band form granulocytes	Neutrophil band form granulocytes	Erythrocytes sedimentation rate	Eosinophil granulocytes
	21	22	11	19	19
Rheumatoid arthritis		Serum calcium	Haemoglobin	Serum uric acid	Serum total cholesterol
		Leukocytes	Neutrophil band form granulocytes	Erythrocytes sedimentation rate	Age
		Serum aspartate aminotransferase	Serum uric acid	Age	Neutrophil band form granulocytes
		33	11	19	21
Lupus erythematosus			Haemoglobin	Erythrocytes sedimentation rate	Serum glucose
			Leukocytes	Serum calcium	Age
			Serum potassium	Age	Serum total protein
			17	19	21
Gout				Haemoglobin	Haemoglobin
				Serum chloride	Serum total cholesterol
				Erythrocytes sedimentation rate	Erythrocytes sedimentation rate
				10	12
Arthrosis					Erythrocytes sedimentation rate
					Age
					Serum albumin
					16

Tab. 6. Comparison between the calculated variables and those found in the literature.

Calculated	Haemoglobin	Leukocytes			Neutrophil band form granulocytes	Serum potassium	Eosinophil granulocytes	Age
Literature	Haemoglobin	Leukocytes	Platelets	Neutrophil segmented granulocytes				Age
Calculated		Serum chloride	Serum uric acid	Serum calcium	Serum inorganic phosphate	Serum glucose	Serum total protein	Serum albumin
Literature	Sex		Serum uric acid			Serum glucose	Serum total protein	Serum albumin
Calculated	Serum total cholesterol	Erythrocytes sedimentation rate			Serum aspartate aminotransferase			
Literature		Erythrocytes sedimentation rate	Blood pressure	Body temperature				

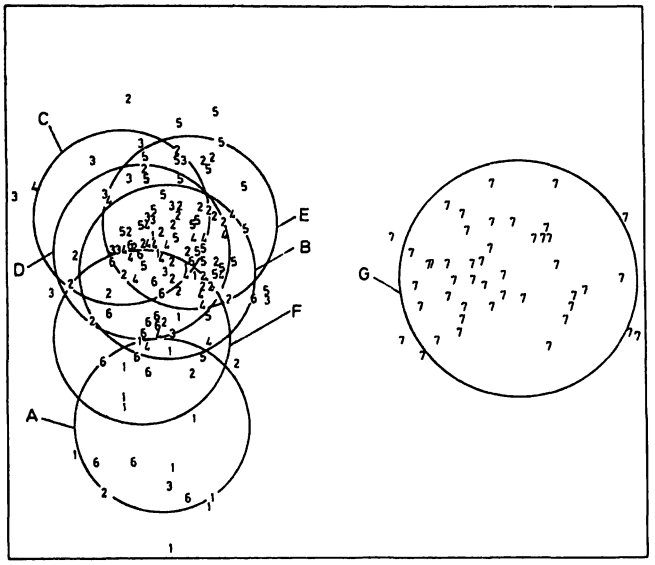

Fig. 1a. Plot of two discriminant scores, resulting from a discriminant analysis on 7 classes and 20 variables:
haemoglobin,
leukocytes,
serum uric acid,
serum calcium,
serum inorganic phosphate,
serum glucose,
serum total protein,
serum albumin,
serum total bilirubin,
serum alkaline phosphatase,
serum lactate dehydrogenase,
serum aspartate aminotransferase,
serum total cholesterol,
erythrocytes sedimentation rate,
neutrophil segmented granulocytes,
neutrophil band form granulocytes,
lymphocytes,
monocytes,
eosinophil granulocytes,
age.
59.6% of the cases were correctly classified. The patients are indicated by a number indicating the disease and the centroid of each group is indicated by a letter:

rheumatic fever	(1,A)
rheumatoid arthritis	(2,B)
lupus erythematosus	(3,C)
gout	(4,D)
arthrosis	(5,E)
osteomyelitus	(6,F)
reference group	(7,G)

By means of a linear discriminant analysis these values are increased to 62% and 51%, respectively. Much better results are gained by reducing the six disease categories to three by combining the infectious diseases, (rheumatic fever + osteomyelitis), the auto immune diseases (rheumatoid arthritis + lupus erythematosus) and 'mechanical' defects (arthrosis + gout) as shown in figure 2 for 21 variables. Table 7 gives the result obtained by linear discriminant analysis. Thus, 71% of the patients can be classified correctly and this score does not alter by reduction of the number of variables to 8.

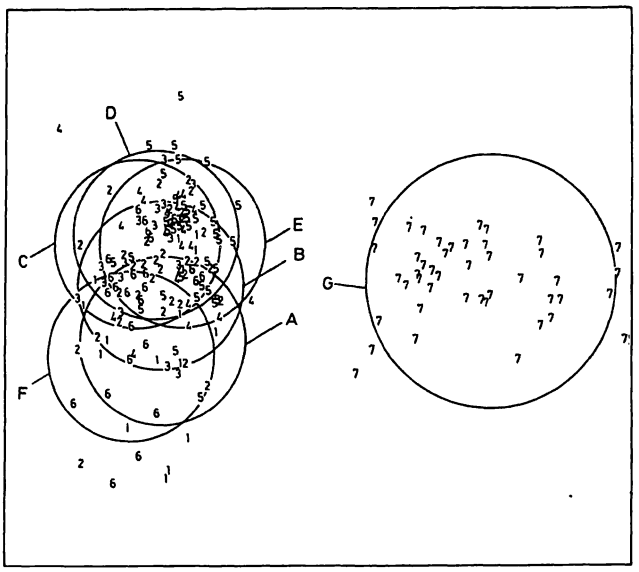

Fig. 1b. Plot of two discriminant scores, resulting from a discriminant analysis on 7 classes and 10 variables:
haemoglobin,
serum uric acid,
serum calcium,
serum inorganic phosphate,
serum albumin,
serum alkaline phosphatase,
serum total cholesterol,
erythrocytes sedimentation rate,
neutrophil band form granulocytes,
age.
57.6% of the cases were correctly classified. The patients are indicated by a number indicating the disease and the centroid of each group is indicated by a letter:

rheumatic fever	(1,A)
rheumatoid arthritis	(2,B)
lupus erythematosus	(3,C)
gout	(4,D)
arthrosis	(5,E)
osteomyelitus	(6,F)
reference group	(7,G)

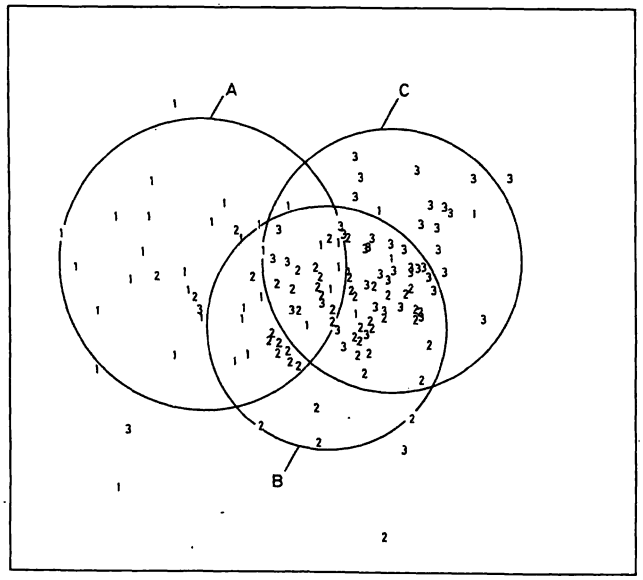

Fig. 2. Plot of two discriminant scores, resulting from a discriminant analysis on 3 classes and 21 variables:

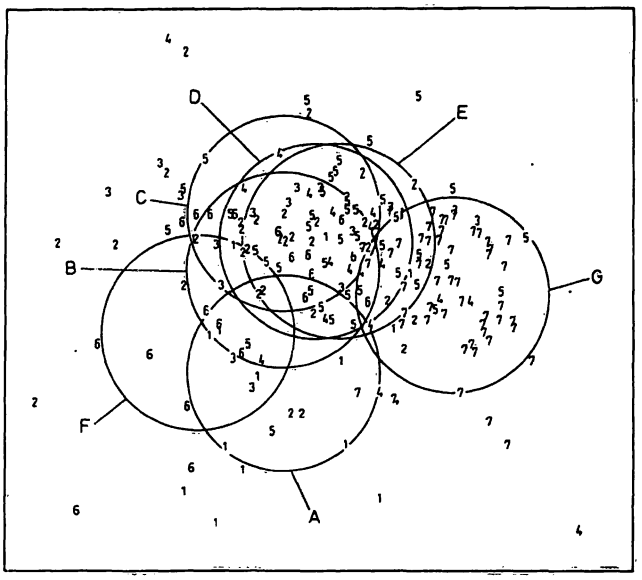

Fig. 1c. Plot of two discriminant scores, resulting from a discriminant analysis on 7 classes and 9 variables (same as in figure 1b minus age).
48.5% of the cases were correctly classified. The patients are indicated by a number indicating the disease and the centroid of each group is indicated by a letter:

rheumatic fever	(1,A)
rheumatoid arthritis	(2,B)
lupus erythematosus	(3,C)
gout	(4,D)
arthrosis	(5,E)
osteomyelitus	(6,F)
reference group	(7,G)

Fig. 2. Continued

serum uric acid,
erythrocytes sedimentation rate,
leukocytes,
serum urea,
serum lactate dehydrogenase,
neutrophil band form granulocytes,
age,
serum total bilirubin,
serum aspartate aminotransferase,
serum albumin,
monocytes,
serum alkaline phosphatase,
serum inorganic phosphate,
serum total protein,
serum glucose,
eosinophil granulocytes,
serum calcium,
haemoglobin,
neutrophil segmented granulocytes,
lymphocytes,
serum total cholesterol.
70.6% of the cases were correctly classified. The patients are indicated by a number indicating the disease class and the class is indicated by a letter:

infections	= rheumatic fever + osteomyelitis (1,A)
auto-immune diseases	= rheumatoid arthritis + lupus erythematosus (2,B)
'mechanical' problems	= arthrosis + gout (3,C)

Tab. 7. Prediction results of the analysis shown in figure 2.

Actual group	Predicted group membership		
	Rheumatic fever + Osteomyelitis	Rheumatoid arthritis + Lupus erythematosus	Arthrosis + Gout
Rheumatic fever + Osteomyelitis	24 (61%)	5 (13%)	10 (26%)
Rheumatoid arthritis + Lupus erythematosus	5 (10%)	34 (64%)	14 (26%)
Arthrosis + Gout	4 (6%)	10 (14%)	57 (80%)

Discussion

The first problem to be studied was the importance of the variable for the discrimination between disease categories. As usual in retrospective studies various data are missing for each patient.

Normally, the resulting problem is solved by substituting the mean value of each variable in the subgroup concerned for the missing data. The importance of each variable in the achieved classification can then be estimated.

However, when there are many missing data, the classification and the estimation become unreliable. Therefore, we applied two different selection methods. Variable selection by two methods with no and partial correlation correction yielded strongly differing results. The weak similarity between the two selections is an expression of the mathematical differences between the two methods but also an indication of dealing with derived, secondary variables that are indicative for a status of being ill rather than for a specific disease.

The two methods together lead to the selection of a set of variables that have a strong resemblance to the variable set selected on clinical grounds. The nature of the selected variables and also the much higher classification score after reduction of the six disease categories to three combined classes clearly prove that the measured chemical and haematological effects accompanying joint diseases are not pathognomic, but probably only the result of an infection or autoimmunity, while the importance of the age and sex related variable once again makes it clear that predisposition to suffer from the disease in question plays a substantial role.

The applied mathematical techniques, as sophisticated as they are, cannot provide more information than is stored in the datasets supplied. Therefore, our final conclusion is that these variables are not relevant to the diagnosis. They may however be helpful in objectivating the complaints and therapy results, since it is clearly possible to separate the disease groups from normal patients (reference group).

Acknowledgement

The authors wish to thank Mrs. *A. J. L. M. Veuger*, M. D., for her helpful cooperation; Prof. Dr. *J. B. J. Soons* and Prof. Dr. *A. Dijkstra* for their encouraging discussions; Prof. Dr. *R. W. Lent* (Albert Einstein College of Medicine, New York) for kindly reviewing the manuscript; Drs. *L. van Norel* for performing some exploratory statistical studies; Miss *A. M. J. van Bommel*, Mrs. *J. W. A. M. van Ingen-van Berkel* and Miss *M. H. van Osch* for their skilful assistance.

References

1. Goldberg, D. M. & Ellis, G. (1978) Adv. Clin. Chem. *20*, 49–128.
2. Wilding, P., Kendall, M. J., Holder, R., Grimes, J. A. & Farr, M. (1975) Clin. Chim. Acta *64*, 185–194.
3. Jørgensen, K. & Astrup, P. (1957) Scand. J. Lab. Invest. *9*, 122–132.
4. Borst, A., Hanssen, C. J. M. & de Jong, E. B. M. (1974) Clin. Chim. Acta *55*, 121–128.
5. Kingsley, G. R. & Getchell, G. (1953) Stand. Meth. Clin. Chem. *1*, 113–117.
6. Harper, A. M., Duewer, D. L. & Kowalski, B. R. (1977) In: Chemometrics, Theory and Practice (Kowalski, B. R., ed.), Am. Chem. Soc. Symp. Ser. Nr. 52, 14–52.
7. Duewer, D. L., Koskinen, J. R. & Kowalski, B. R. (1975) "ARTHUR" (available from B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry BG-10, University of Washington, Seattle, Washington 98 195).

8. Solberg, H. E. (1978) Discriminant Analysis in Clinical Chemistry. CRC. Crit. Rev. Clin. Lab. Sci., 209.
9. Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K. & Bent, D. H. (1975) Statistical Package for the Social Sciences (SPSS), McGraw-Hill, New York, 2nd ed., 434–467.
10. Holvex, D. N. (ed.) & Talbott, J. H. (cons. ed.) (1972) The Merck Manual, 12th ed., Merck Sharp and Dohme Research Laboratories, Rahway.
11. Eastman, R. D. (1975) Biochemical values in clinical medicine: The results following pathological or physiological change, Wright Ltd., Bristol.
12. Bondy, P. K. & Rosenberg, L. E. (1980) Metabolic control and disease, 8th ed., W. B. Saunders Company, Philadelphia, London, Toronto.

Drs. H. M. J. Goldschmidt
Director
Department of Clinical Chemistry
and Haematology
Maria Ziekenhuis
Dr. Deelenlaan 5
Postbus 90107
NL-5042 AD Tilburg